Outline

• Recap

• Tumor Phylogeny Inference from Bulk DNA-seq with Copy-Number Aberrations

Reading:

Tumor Evolution as a Phylogenetic Tree

**Clonal Theory** [Nowell, 1976]

- **Normal cell**
- **Founder clone**
- **Somatic mutation**

Present time → Normal cell → Founder clone → Somatic mutation → Tumor → Phylogenetic Tree
Observations are Leaves of a Perfect Phylogeny $T$

Assumptions:
- Mutations are single nucleotide variants (SNVs)
- Infinite sites assumption

Single-cell sequencing

Tumor Snapshot

$M = \begin{bmatrix}
1 & 1 & 1 & 0 & 0 & 0 \\
1 & 1 & 0 & 1 & 0 & 0 \\
1 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & 0 & 1 & 0 & 0 \\
1 & 0 & 0 & 0 & 1 & 1 \\
\end{bmatrix}$

Binary Matrix $B$

Two-State Perfect Phylogeny Tree $T$

$O(mn)$

[Gusfield, 1991]

Seq. method | Inferring $T$ | Complexity
--- | --- | ---
single-cell | unmixed two-state perfect phylogeny | $O(mn)$
Observations are Mixtures of the Leaves of $T$

Bulk sequencing

Variant Allele Frequencies Matrix $F$

SNVs

\[
F = \begin{bmatrix}
0.4 & 0.0 & 0.0 & 0.0 & 0.3 & 0.2 \\
0.3 & 0.3 & 0.0 & 0.3 & 0.0 & 0.0 \\
0.4 & 0.4 & 0.4 & 0.0 & 0.0 & 0.0
\end{bmatrix}
\]

Two-State Perfect Phylogeny Tree $T$

Mixing Proportions $U$

Seq. method & Inferring $T$ & Complexity
--- & --- & ---
single-cell & unmixed two-state perfect phylogeny & $O(mn)$
bulk & mixed two-state perfect phylogeny & NP-complete

TrAp [Strino et al., 2013], PhyloSub [Jiao et al., 2014], CITUP [Malikic et al., 2015], BitPhylogeny [Yuan et al., 2015], LICHeE [Popic et al., 2015], ...
Copy-Number Aberrations Confound VAFs

Need > 2 states:
- 0: non-mutated
- 1: mutated
- 2: single-copy amplification
- 3: ...

VAF: 0.4

SNVs
SNVs

F =

\[
\begin{bmatrix}
0.4 & 0.0 & 0.0 & 0.3 & 0.2 \\
0.3 & 0.4 & 0.0 & 0.3 & 0.0 \\
0.4 & 0.4 & 0.4 & 0.0 & 0.4
\end{bmatrix}
\]
Outline

• Multi-State Perfect Phylogeny Mixture Problem

• Combinatorial Characterization of Solutions

• Application to Cancer Bulk-Sequencing Data

• Results
Infinite Sites Generalizes to Infinite Alleles

**Two-State Perfect Phylogeny** – Infinite sites assumption: a character changes state once

Frequency Matrix $F$

$$
\begin{array}{c c c c c c c}
S_1 & S_2 & S_3 \\
0.8 & 0.8 & 0.8 & 0.0 & 0.0 & 0.0 & 0.0 \\
0.6 & 0.6 & 0.0 & 0.6 & 0.0 & 0.0 & 0.0 \\
0.8 & 0.0 & 0.0 & 0.0 & 0.6 & 0.4 &
\end{array}
$$

$n$ characters

$$
\begin{pmatrix}
1 & 0 & 0 & 0 & 0 & 0 \\
1 & 1 & 0 & 0 & 0 & 0 \\
1 & 1 & 1 & 0 & 0 & 0 \\
1 & 1 & 0 & 1 & 0 & 0 \\
1 & 0 & 0 & 0 & 1 & 0 \\
1 & 0 & 0 & 0 & 1 & 1 \\
\end{pmatrix}
$$

Two-State PP Matrix $B$

Two-State PP Tree $T$
Infinite Sites Generalizes to Infinite Alleles

**Two-State Perfect Phylogeny** – Infinite sites assumption: a character changes state once

VAF Factorization Problem (VAFFP): Given $F$, find $U$ and $B$ such that $F = UB$ [El-Kebir, Oesper et al., 2015]
Infinite Sites Generalizes to Infinite Alleles

**Two-State Perfect Phylogeny** – Infinite sites assumption: a character changes state once

\[
F = U B
\]

[El-Kebir, Oesper et al., 2015]

**VAF Factorization Problem (VAFPP):** Given \( F \), find \( U \) and \( B \) such that \( F = U B \)

**Multi-State Perfect Phylogeny** – Infinite alleles assumption: a character changes to a state once

\[
\begin{pmatrix}
0 & 0 \\
1 & 1 \\
2 & 0 \\
0 & 1 \\
0 & 2 \\
\end{pmatrix}
\]

Multi-State PP Matrix \( A \)

Multi-State Perfect Phylogeny Tree \( T \)
Infinite Sites Generalizes to Infinite Alleles

**Two-State Perfect Phylogeny** – Infinite sites assumption: a character changes state once

\[
F = U B
\]

Two-state Perfect Phylogeny Mixture Problem: Given \(F\), find \(U\) and \(B\) such that \(F = U B\) [El-Kebir, Oesper et al., 2015]

**Multi-State Perfect Phylogeny** – Infinite alleles assumption: a character changes to a state once

\[
F_0 U = A_0 U
\]

\[
F_1 U = A_1 U
\]

\[
F_2 U = A_2 U
\]

Multi-state Perfect Phylogeny Mixture Problem: Given \(F\), find \(U\) and \(A\) such that \(F_i = U A_i\) for all states \(i\)

Multi-State Perfect Phylogeny Tree \(T\)
Combinatorial Characterization of Solutions

Two-State Perfect Phylogeny

- **m samples**
  - $(0.8, 0.8, 0.8, 0.0, 0.0, 0.0)$
  - $(0.6, 0.6, 0.0, 0.6, 0.0, 0.0)$
  - $(0.8, 0.0, 0.0, 0.0, 0.6, 0.4)$

- **Frequency Matrix $F$**

- **Ancestry Graph $G$**
  - Directed acyclic graph
  - Potential parental relationship

- **Theorem** [El-Kebir, Oesper et al., 2015; Popic et al., 2015]
  Solutions are **spanning trees** that satisfy
  \[
  f_{p,(c,1)} \geq \sum_{(d,1) \in \delta(c,1)} f_{p,(d,1)} \quad (SC)
  \]

Multi-State Perfect Phylogeny

- **m samples**
- **n characters**
- **Frequency Tensor $\mathcal{F}$**

**Theorem** [El-Kebir, Oesper et al., 2015]
VAFPP is NP-complete for $m = O(n)$
Combinatorial Characterization of Solutions

**Two-State Perfect Phylogeny**

- **Frequency Matrix** $F$
- **Ancestry Graph** $G$
- **Theorem** [El-Kebir et al., 2016]
  - PPMDP is NP-complete even for $m = 2$ and $k = 2$

**Multi-State Perfect Phylogeny**

- **Frequency Tensor** $\mathcal{F}$
- **Multi-State Ancestry Graph** $G$
- **Theorem** [El-Kebir, Oesper et al., 2015]
  - VAFFP is NP-complete for $m = O(n)$
- **Theorem** [El-Kebir et al., 2016]
  - Solutions are threaded spanning trees satisfying
    \[
    f_p^+(D(c,i)) \geq \sum_{(d,j) \in \delta(c,i)} f_p^+(D(d,j)) \quad (\text{MSSC})
    \]

**Ancestry Graph** $G$

- Directed acyclic graph
- Mutations / characters
- Potential parental relationship

**Character-State Pair**

- Directed multi-graph
- Potential parental relationship labeled by descendant sets

**Combinatorial Characterization of Solutions**
Outline

• Multi-State Perfect Phylogeny Mixture Deconvolution Problem

• Combinatorial Characterization of Solutions

• Application to Cancer Bulk-sequencing Data

• Results
Application to Cancer Bulk Sequencing

Frequency Tensor $\mathcal{F}$

- $m$ samples
- $n$ characters
- $q$ copies
- $p$ copies
- $d$ copies
- $c$ copies
- $z$ copies

$F(x, y, z)$: Frequency Tensor

State Graph

- $x$: # maternal copies
- $y$: # paternal copies
- $z$: # mutated copies

State Graph:

- $(1, 1, 0)$
- $(1, 1, 1)$
- $(2, 1, 2)$

Equations:

- $m = (1, 1, 0)$
- $n = (1, 1, 1)$
- $q = (2, 1, 2)$

Maternal copies:

- $(x, y, z)$

Paternal copies:

- $(x, y, z)$

Mutated copies:

- $(x, y, z)$
Application to Cancer Bulk Sequencing

Frequency Tensor $\mathcal{F}$

- # maternal copies
- # mutated copies
- # paternal copies

$(x, y, z)$

State Graph

- $x$: # maternal copies
- $y$: # paternal copies
- $z$: # mutated copies

Copy-number calls: THetA, Battenberg, ...

VAF

- $\mu_{(1,1)} = 0.8$
- $\mu_{(2,1)} = 0.2$
- $h = 3/10 = 0.3$

State Trees

- $S_1$
- $S_2$
- $S_3$
- $S_4$

CCFs

State Trees

✓✗✓✓

$\implies$

- $x = (1, 1, 0)$
- $y = (1, 1, 1)$
- $z = (2, 1, 2)$
SPRUCE Enumerates Phylogenies

**Somatic Phylogeny Reconstruction Using Combinatorial Enumeration**
Available at: http://compbio.cs.brown.edu/projects/spruce/

Copy-number calls and mixing proportions

\[
\begin{align*}
\mu_{(1,1)} &= 0.8 \\
\mu_{(2,1)} &= 0.2
\end{align*}
\]

VAF

\[
\begin{align*}
&\text{variant reads} \\
&\text{reference reads}
\end{align*}
\]

\[
h = \frac{3}{10} = 0.3
\]

VAF confidence interval

\[
[h, \overline{h}] = [0.18, 0.32]
\]

Error-Model for Variant Allele Frequencies
SPRUCE:
Somatic Phylogeny Reconstruction Using Combinatorial Enumeration
Outline

• Multi-State Perfect Phylogeny Mixture Deconvolution Problem

• Combinatorial Characterization of Solutions

• Application to Cancer Sequencing

• Results
SPRUCE Accurately Recovers Simulated Trees

Parameters:
• # characters $n$: 5, 15
• # samples $m$: 2, 5, 10

Methods:
• SPRUCE
• PhyloWGS [Deshwar et al., 2015]

Increasing number of samples decreases ambiguity
SPRUCE Accurately Recovers Simulated Trees

Parameters:
• # characters $n$: 5, 15
• # samples $m$: 2, 5, 10

Methods:
• SPRUCE
• PhyloWGS [Deshwar et al., 2015]

Increasing number of samples decreases ambiguity
SPRUCE Accurately Recovers Simulated Trees

Parameters:
- # characters \( n \): 5, 15
- # samples \( m \): 2, 5, 10

Methods:
- SPRUCE
- PhyloWGS [Deshwar et al., 2015]

Increasing number of samples decreases ambiguity
More Samples vs. More Coverage

- Sequencing depth
- Error-free
- # Samples
- # Solutions
- n = 5 characters

- # Solutions
- m
- 2
- 5
- 10
Violations of Infinite Alleles Assumption

The diagram illustrates the recall distribution for different numbers of violations (0, 1, 2, 3, 4, 5). Each box plot represents the distribution of recall values for a given number of violations, with the median and interquartile range shown. The x-axis indicates the number of violations, while the y-axis represents the recall values.
Cancer Cell Fractions (CCFs) Cannot Be Inferred A Priori

CCFs are used extensively in studying intra-tumor heterogeneity and tumor evolution:

1. Timing of driver mutations

2. Tumor evolution and phylogeny reconstruction

3. Developmental patterns of metastases

\[
\text{CCF}(\bigotimes) = \frac{\text{# tumor cells with } \bigotimes}{\text{# tumor cells}}
\]
Metastatic Evolution in Prostate Tumor

**Input:**
- 10 samples: whole-genome & targeted sequencing
- ~110 SNVs

**Tree building:**
1. Infer cancer cell fraction (CCF) for each SNV in each sample
2. Cluster SNVs by CCFs across samples
3. Construct tree using Pigeon-Hole-Principle (Sum Condition)

Cancer Cell Fractions Cannot Be Inferred A Priori

Compatible state trees and state frequencies

PPFIA1 Cancer Cell Fraction

Number of samples with distinct CCF intervals

TAS2R4
ENSG0000196960
GRIN2B
CCDC21
SENP6

PPFIA1

Cancer Cell Fractions Cannot Be Inferred A Priori

Compatible state trees and state frequencies

PPFIA1 Cancer Cell Fraction

Number of samples with distinct CCF intervals

TAS2R4
ENSG0000196960
GRIN2B
CCDC21
SENP6

PPFIA1

Cancer Cell Fractions Cannot Be Inferred A Priori

Compatible state trees and state frequencies

PPFIA1 Cancer Cell Fraction

Number of samples with distinct CCF intervals

TAS2R4
ENSG0000196960
GRIN2B
CCDC21
SENP6

PPFIA1

Cancer Cell Fractions Cannot Be Inferred A Priori

Compatible state trees and state frequencies

PPFIA1 Cancer Cell Fraction

Number of samples with distinct CCF intervals

TAS2R4
ENSG0000196960
GRIN2B
CCDC21
SENP6

PPFIA1

Cancer Cell Fractions Cannot Be Inferred A Priori

Compatible state trees and state frequencies

PPFIA1 Cancer Cell Fraction

Number of samples with distinct CCF intervals

TAS2R4
ENSG0000196960
GRIN2B
CCDC21
SENP6

PPFIA1

Cancer Cell Fractions Cannot Be Inferred A Priori

Compatible state trees and state frequencies

PPFIA1 Cancer Cell Fraction

Number of samples with distinct CCF intervals

TAS2R4
ENSG0000196960
GRIN2B
CCDC21
SENP6

PPFIA1
Conclusions

- Copy-number aberrations confound variant allele frequencies
  - SNVs and CNAs must be considered jointly in phylogeny reconstruction

- Generalization of infinite sites for SNVs is infinite alleles for SNVs + CNAs
  - Multi-state Perfect Phylogeny Mixture Problem (PPM)

- Complete combinatorial characterization of the problem
  - Solutions are constrained spanning trees in a directed multi-graph
  - PPM is NP-complete for $k = 2$ and $m = 2$

- Using combinatorial structure, SPRUCE accurately recovers simulated trees

- Cancer cell fractions cannot be uniquely inferred \textit{a priori} by considering SNVs in isolation

- Precise mathematical models are needed to describe evolutionary process in cancer